Googling the news

Opportunities and challenges in studying news events through Google Search

Ørmen, Jacob

Published in:
Digital Journalism

DOI:
10.1080/21670811.2015.1093272

Publication date:
2016

Document version
Early version, also known as pre-print

Googling the news: Opportunities and challenges in studying news events through Google Search

Pre-print version of Ørmen, J. Googling the news: Opportunities and challenges in studying news events through Google Search (2016). *Digital Journalism, 4(1):* http://dx.doi.org/10.1080/21670811.2015.1093272

Abstract:
Search engines provide a window into the changing association between websites and keywords across cultures and countries and over time. As such, they offer journalism and news researchers an opportunity to study how search engines, in this case Google, mediate news events and stories online. However, search results are not straightforward to study. Since search results are made in the act of searching and will have to be retrieved from Google Search in real-time, there is a range of different ontological and methodological issues related to this data source. This paper addresses these issues by discussing how factors in the search algorithm can be used proactively to study variations across searchers and in time. The paper identifies various endogenous and exogenous factors in the search algorithm one has to pay attention to and discusses ways to archive search results accordingly. Through a small case study, ways to work with the influence of endogenous factors (keywords, language settings, geo-location, web history and clicking behaviour) and mitigate the effects of the exogenous factors (experimentation and randomization) are suggested. Then, a new approach to studying search results is put forward, which builds on purposeful sampling of real-world participants or constructed research profiles. Finally, perspectives for news and journalism scholars in studying algorithmically generated content in a broader context are offered.

Keywords: Google, search results, algorithm, personalisation, news event,

Introduction

Search engines in general and Google Search in particular remain important entry points to the web for the majority of people in the Western world – approximately 65% use Google regularly in the US and probably more than 95% in Europe (Hillis, Petit, & Jarrett, 2013). Accordingly, search engines function as important gatekeepers (Bozdag 2013) for people to find information about topics, major events, news stories, disasters, etc.; and this provides us a "unique empirical window into the study of culture" (Sanz and Stancik 2013). Since most people only click on results at the top of the search page (Pan et al. 2007), the constellation of search results (commonly referred to as the search rankings) are of great importance in determining which perspectives, angles, and topics on a given event or story are most salient for the public eye. During big news events, such as political elections, the importance of search engines increases as people turn to the web to find more information about what is going on. For instance, during the 2015 UK General Election, search interest on Google for the leading candidates David Cameron (Conservative) and Ed Miliband (Labour) were greater than ever before – more than in previous elections and 5-10 times the number compared to non-election months[i]. For news organisations, search engines remain a key traffic source to stories online – on par with or, in some cases, surpassing social media (Mitchell, Jurkowitz, and Olmstead 2014). Thus, it is important to document and study how search results appear across time and space to understand how Google functions as a gatekeeper or mediator of news and information to the public as well as online traffic to news organisations

However, studying search results is anything but straightforward. First of all, search engines generally do not come with ready-made ways to retrieve data. Whereas social network sites such as Facebook and Twitter have become a popular way to study the online distribution of news stories and events (see Vis 2013; Bruns et al. 2012; Bro and Wallberg 2014) because of the relatively easy access they provide to data through Application Programming Interfaces (API), Google does not permit retrieval of the rankings for specific keywords. Secondly, in contrast to offline documents, online articles, posts or tweets, search results do not simply exist 'out there' waiting to be found and analysed but have to be created in the act of searching. They exist as a particular type of document online (algorithmically-generated content), based on one hand, in the index of retrievable documents on the web that

the search engine compiles, and on the other hand, guided by a range of factors in the search algorithm that is informed by the person searching – *intentionally* (e.g., through search keywords) as well as *unintentionally* (e.g. trough personalization). In that sense, the particular results are a co-creation of the person searching (through signals) and the search engine (by providing an index to search). They are, so to speak, not "data found" by the researcher but "data made" in the research interaction (Jensen 2010). They exist as the unique coming-together of an individual searcher and the index as it existed at that particular moment. This entails that they cannot be retroactively retrieved or reproduced. Therefore, we need systematic approaches to the documentation of search results in real time that can deal with these ontological and methodological challenges.

In this article, I present a first step in establishing a methodology for search engine analysis by discussing the opportunities and challenges in studying news events through Google Search. I use *events* here instead of just *stories* or *coverage* to emphasise the pre-planned, recurring or, at least, expected manner of occurrence. This understanding of events draws on the familiar concept of "media events" (Dayan & Katz 1992) but without the stringent semantic, syntactic, and pragmatic genre requirements discussed in the original studies and with a narrower focus on events covered by the news – rather than all – media. The reason for limiting the study to pre-planned events is that they are much easier to deal with when setting up a research design for documenting search results. Since these events comprise only a subset of news stories that it could be of interest to document through Google Search, I will address the difficulties of studying search results in relation to unexpected and disruptive events such as natural disasters and terrorism.

The article begins with a discussion of how factors in the search algorithm seem to influence search results and how this affects the way we can reasonably study search engines. Then, it turns to a discussion of feasible ways to archive these results and how research projects can actively work with some factors and seek to mitigate others. In this process, it provides examples from a small case study of a particular news event - Felix Baumgartner's skydive from the stratosphere in 2012, called the Red Bull Stratos. Finally, strategies for how to work with – instead of against – the algorithms with real-world participants or programmed research profiles are assessed.

Factors in the algorithm: Opening the black box of uncertainty

The greatest challenge facing research into search results is that the logics of search rankings are very difficult to uncover. The number of factors, all the signals the search algorithm takes into account when compiling the rankings for the individual search query, informing the search rankings, and the individual weight of each factor are impossible for outsiders not affiliated with Google to uncover fully. A number of studies have tried to 'second-guess' Google's search algorithm(s) through systematic, large-scale mapping of search rankings across queries (see, e.g., Edelman 2010; Edelman and Lockwood 2011; Jiang 2012); but, in recent years, it has become increasingly clear that the multitude of factors that inform the exact constellation of search results for any given query (Granka 2010) as well as the increasing adaptation of results to individual users – personalisation – have made this task very difficult (Feuz, Fuller, and Stalder 2011). Because of the 'black-boxed' (Marres and Weltevrede 2013) functions of these search algorithms, there simply exists no vantage point from which researchers can analyse search results objectively. Reverse-engineering the search algorithm seems neither technically feasible nor scientifically desirable. This does not mean, however, that it is not interesting or relevant to consider the workings of the search algorithm when studying search results.

Traditionally, search engines have operated on *query-dependent* factors - comprising the position and order of search queries, the amount and types of relevant keywords in the web documents as well as language settings and the geo-location of the user - and *query-independent* factors - including, among other things, the popularity of web sites – determined, for instance, by the PageRank algorithm (Brin & Page 1998), the freshness of the documents, and click popularity (Lewandowski 2005). Recently, Google has announced that it will start to display mobile-ready sites more prominently in searches conducted through mobile devices (Makino, Jung, and Phan 2015). These factors can also be supplemented with more general maintenance tasks done by the search operator (e.g., Google). This includes randomisation and experimentation (Zuckerman 2011). When Google receives a number of queries from the same IP address, it might deliberately try to randomise search results a bit in order to mask the workings of the algorithm (Zuckerman 2011). At the same time, Google is constantly conducting experiments (e.g., A/B tests) to detect the kind of search results (and design

elements) users are most likely to interact with (Zuckerman 2011). Thus, randomisation and experimentation can easily introduce random error into the search result study on top of the query-dependent and –independent factors.

In recent years, a third type has increasingly come to inform search results: *personalisation* factors. There is no clear-cut definition of what personalisation encompasses, but it includes all those signals search engines might use to adapt the search results to the individual user – for example, prior search history, browsing behaviour, and whether the user is logged into services (e.g., Google Accounts). Even though the knowledge and criticism of personalisation measures have been around for many years (Zimmer 2008; Spink et al. 2003), only recently has personalisation been empirically tested by research. One study used artificial research profiles with designed browser histories to compare personalisation across a number of queries (Feuz, Fuller, and Stalder 2011). These researchers find personalisation to be extensive but did not attribute it to specific factors, and they held geo-location as a constant. A recent study of real-world search participants finds that, on average, about 11.7% of Google search results are personalised (Hannak et al. 2013) , which would suggest that personalisation is not as influential as previously assumed (*pace* Pariser 2012). The authors do note, however, that there is considerable variation across topics - 'politics' and 'news' showcase greater levels of personalisation and tend to fluctuate more (Hannak et al. 2013). Hannak and colleagues find geo-location based on users' IP addresses and Google Accounts login to be the sole significant causes of personalisation (ibid). A different study using a similar design finds variation across about 98% of the results (Xing et al. 2014) but also identifies geo-location to have the greatest influence. Thus, personalisation can be considered as the great 'known unknown' of search engines – we know it is there but not exactly how it works – that seems to exercise lesser or greater influence on the rankings, depending on the type of search and the steps taken by users to avoid personalisation.

In short, the number and type of factors that go into the search algorithm are very difficult to assess. The factors discussed here are merely those commonly identified in the literature, which is an ever-open list that remains largely unknown. Nonetheless, I find it helpful to divide the range of factors discussed above into two analytical categories: One comprised of all the factors that the individual searcher and, thus, researchers have some control of, and another consisting of all the factors that are beyond our reach when using the search engine. I call the first category: *endogenous factors*, since these factors are, so to speak,

endogenous to the particular act of searching. It includes the search queries, language settings, geo-location as well as the known list of personalisation factors (search history, browsing behaviour, whether one is logged into services). The second category, *exogenous factors*, includes all those factors that are beyond the direct control of the user and affect the search results on a more general level. This involves experimentation and randomisation done by Google as well as all the factors that go into deciding the PageRank and other measures of importance for web sites.

The primary research interest is in the relationship between the endogenous factors on the one hand and the exogenous factors on the other. It is foremost the exogenous factors dealing with rankings of websites based on importance that is of primary interest, since we would assume experimentation and randomization to be affecting the search results more by chance, or at least not purposefully. The latter can be considered akin to random error that one has to watch out for in the study. To study this relationship between endogenous and exogenous factors we need archiving methods that can take the workings of search algorithms into account.

Retrieving and archiving rankings for search results

The practical question of how to archive search results has only become more complicated in recent years. Whereas the popularity of individual keywords across time can be accessed through Google's own tool Google Trends, there is currently no interface that can retrieve the exact constellation of websites (the search results) for these keywords. The official API to search through the whole index of Google Search has been discontinued by Google as of November 2010 and replaced by a Custom Search API that offers very limited search options.[ii] It also appears that the results produced by the search APIs (both the present one and the discontinued one) produce quite different search results from manual searches (Hannak et al. 2013). This again reiterates the lack of a baseline search result ranking to which to compare individual searches. This entails that results will have to be scraped or retrieved in other ways directly from the search page (e.g., google.com or affiliated subdomains) by individual searchers themselves or through software.

The first issue with this is how to produce and archive the search results for later analysis. One commonly applied method for retrieving search results is to use programs that can query Google automatically and repeatedly. These programs can be designed from the bottom up through various programming languages, or one can use read-made programs, e.g., Google Search Scraper by Amit Agarwal (Agarwal 2015). The upside of this approach is that it is possible to set up "lobster traps" that sit passively and collect data, waiting for interesting changes to happen (Karpf 2012). However, this solution is problematic for four reasons. First, Google bans this option in the Terms of Service (ToS)[iii] and it is generally seen as a 'dirty research method' (Rogers 2013a). Second, since the organic search results are at the core of Google's business model, it is considered proprietary and guarded with great care. Therefore, Google is particularly aware of attempts to scrape search results repeatedly, so there is a high risk of facing restrictions on access (e.g., through CAPTCHAs) or getting the IP banned for a shorter or longer amount of time. Figure 1 shows an example of the latter. Third, ironically, these programs easily introduce a layer of complexity on top of the search algorithm. When using a piece of software that retrieves the search results automatically, it can be hard to tell or adjust the signals provided to the search engine by the program (e.g., the language settings, IP address, and web history). Fourth, the scraped data usually only includes the organic search results and, thus, excludes the paid-for search content (usually, on top of and to the right of the organic content) as well as other information displayed in the search window (such as fact boxes and other information displayed by Google on the search result page).

[FIGURE 1 ABOUT HERE]

To retain the visual information in the search results, one can use research tools that can make automated screen dumps of search results at regular intervals. This approach has the benefit of retaining all the visual information from the browser window (Karlsson and Strömbäck 2010) that might work well in research projects that are more interested in the constellation of search results (e.g., the size of each result in the query list, the placement of links, images, videos, and other contextual data). It has its obvious drawbacks if the goal is to conduct statistical analysis, since the information is 'flattened' in one image instead of being nicely ordered in a structured database. Manual recoding of features in the image into quantifiable variables is, of course, possible (Kautsky and Widholm 2008) but will quickly

become quite laborious. Therefore, screen dumping functions work best in small-n studies that integrate the visual elements in the analysis.

A second option would be to let the archiving be done by real-world searchers (participants in the study), e.g., as manual screen dumping. This form of micro-archiving (Brügger 2011) has the great advantage that it stays closest to real acts of searching without the risk of being too artificial like the programmed scraping. The real-world searchers would be instructed to search for specific keywords and take a screenshot every time they make searches. The downside of this approach is that one relies on people actually completing the tasks assigned to them and that the researcher has no direct influence on how people do the search (for instance, whether they choose to be logged in, clean their browser cache, etc.). In short, there is no superior method for archiving search results. One is technically difficult to set up and manage properly (programmed archiving) and the other is costly and relies on compliance by others (human archiving). It depends on the type of archiving one wish to conduct.

A second issue is how much to archive. Since the search results will have to be generated and saved in real-time, this depends on the scale of the research project. Niels Brügger has outlined three different strategies for archiving websites more generally: *snapshot* (a large range of websites at one point in time), *selective* (a narrow list of important sites for prolonged periods), and *event* (archiving websites particularly relevant for a specific event) strategies (Brügger 2011). I introduce three models for search result archiving here that correspond closely to Brügger's strategies:

1. *The cross-sectional model:* The most basic form of search result archiving would be to have multiple real-world searchers or computer programs conduct searches for specified keywords *at one point in time*. Then, variations could be assessed according to the sampling of real-world searchers (variations across geographic location, languages, gender, age, etc.) or the settings for the computer programs, that is, the construction of search profiles according to criteria mimicking real-world searchers. This model is suitable for suddenly occurring stories (breaking news, memes, etc.) where there is no time to plan ahead. This model would make a strong case for establishing similarity in search results, given different searchers and profiles, e.g. the same major news outlets on top of the search results across all or most of searches.

The drawback with a cross-sectional model would be that it would be hard to assess whether large variations are due to the characteristics of the searchers or profiles or due to randomness, experimentation, or a range of undisclosed factors in the algorithm. This would require multiple points of observation to assess.

2. *The longitudinal model*: Another approach would be to archive search results for specific queries *at multiple points in* time, e.g., on a regular pre-planned basis such as once per week for several years in a row. The point here would be to document certain keywords that retain salience in the popular mind (politicians, societal institutions, etc.). This approach, first of all, makes it possible to engage with how search results of specific terms appear in various stages over time (Borra and König 2013) and in various geographic locations. In that way, we can conduct "studies at the micro-level" (Granka 2010) as a supplement to general search trends at the macro-level (as offered, for instance, by Google Trends).

3. *The short burst model*: Another approach would be to collect more material in shorter, yet more intensive waves of documentation. This model is especially relevant for documenting the developments of news events, since high publication frequency by news outlets around spectacular stories tends to cause fluctuations in the search rankings (Hannak et al. 2013) and search activity increases around big events (as mentioned in the introduction). Here, the objective could be to investigate the various news sources that attain the highest rankings in various geographical regions and over time. By relating the results to online and offline news media coverage and social network sites activity, it is possible to compare the relative importance and prevalence given to certain angles, perspectives, news organisations, etc. Furthermore, by archiving search results before, during, and after certain influential events have occurred (e.g., major pre-planned spectacles such as elections), we could assess how these events influence the relative ranking of search results. The short burst can still provide an interesting insight into the shifting constellations of search results during important news events.

Naturally, it is possible to combine these models in a hybrid (for an example in a different context, see the Danish web archive[iv]), and the models outlined above should be seen more as archetypical approaches to the archiving of search results than precise recipes. All three models require us to attend to the ways the search algorithm work and possibly affect the searches. As mentioned earlier, news events (short burst model) offer us a particularly good case for studying how search results vary across searchers. Below, I explore various ways to work with the endogenous and exogenous factors in relation to a specific case, The Red Bull Stratos.

Working with the algorithm: The Case of the Red Bull Stratos

On 14 October 2012 at about 12:08 MDT, some 38 kilometres above the face of the earth, Felix Baumgartner (an Austrian skydiver) stepped out of a capsule and jumped into the stratosphere, beginning his 4-minute-long record-breaking freefall towards the ground. Millions followed the event (named the 'Red Bull Stratos' after its main sponsor) through simultaneous live streams on the web (YouTube alone reported more than 7 million viewers at its peak moments) and on Discovery Channel (which obtained the highest ratings for a non-prime-time programme ever) (Heitner 2012). Most importantly in this context, it was a pre-planned event that could be documented using the short burst model introduced above. The goal was to document how dominant news organisations would be in the search rankings before, during, and after the event. The project was also interested in seeing whether there would be any variations in search rankings in three countries: Denmark (where the researcher is based), the US (where the jump took place), and Austria (where the jumper is from). Here, I only use the event to illustrate how the factors can be worked with, not to try to answer these questions directly.

A screen dumping method was chosen to archive the search results. This was largely because the initial research design required the full content of the search result page as it appeared to human searchers, that is, with organic and paid searches, as well as video and images shown. The tool Siteshoter (only for PC) was chosen, because of its reliability and the ability to make automated screen dumps of websites at specified intervals. Google Results could be retrieved by posting the full URL to the specific query, e.g.,

https://www.google.com/search?q=felix+baumgartner. Siteshoter can only capture the first page of search results with the predefined setting in Google of 10 search results per page. Since it was not clear how fast the rankings would change, a decision was made to document the event every hour from four days before until two days after the jump was scheduled to take place. In total, about 600 screen dumps were made of queries on Google.com, Google.dk (Danish sub-domain), and Google.at (Austrian sub-domain).

Obviously, when one tries to document an event like this through Google Search, the exact keywords used as search queries are of the greatest importance, since they determine the exact angle taken on the subject.

Endogenous factors: keywords

Finding the right keywords poses similar problems to identifying hashtags for Twitter studies. It is very difficult to know in advance which search queries (keywords or phrases) will resonate with a broader population. Like hashtags, one has to pay attention to what is hot and what is not among real-life searchers. Luckily, Google provides a tool to do exactly that, Google Trends. This tool allows for comparisons of search intensity for various search queries over time and across geographical space. Thus, it is possible to find out which of several keywords have been used most intensively by searches at various points in time and in different regions of the world. The past is, of course, only indicative of future behaviour to a certain extent; memes, political slogans, sudden events, etc., will attain sudden and often unexpected attention from a large number of searchers. On those occasions, one would have to rely on keywords popular in news or social media and, then, make qualified guesses. Therefore, Google Trends is most helpful if one seeks to document predictable or recurring events such as the Red Bull Stratos.

The decision to archive searches in relation to the Red Bull Stratos was made about a week before the event took place. At that point, both the event itself and the main character, Felix Baumgartner, the jumper, had already received lot of attention from established news media. To find the exact queries for the study, a number of different words were tested on Google Search in the days leading up to the jump. After comparing queries affiliated with event with the queries affiliated with the jumper, it was quite obvious that they captured very different aspects of the event. The event queries associated the event much more with the

official sources (incl. Red Bull itself), whereas a search for the jumper yielded more person-focused results (among other things, his Facebook page). Eventually, two queries were chosen for the study: one capturing the event itself, 'red bull stratos', and one for the jumper, 'Felix Baumgartner' More general queries such as 'jump', 'stratos', 'felix' & 'baumgartner' were tested but proved to include many search results not specifically relevant for the purpose of mapping this particular event. The decision was made to stick with the more precise queries rather than a catch-all approach (obviously, many people looking for information about the event would use different search terms). Faced with this issue, I decided that false negatives (excluding relevant results) were a better option than false positives (including too many irrelevant results). Making decisions like this depends on the context of the study, naturally.

Google Trends also came in handy in the selection process. After comparing the search volume on Google Trends for the two queries, it was clear that "Felix Baumgartner" had a greater resonance than "Red Bull Stratos" in the total population of searchers.[v] It was also apparent that the popularity of the search queries was greatest in the Central European countries – particularly, Austria, even though the event took place in air space above Nevada in the US. Since Felix Baumgartner is Austrian, the fact that Austrians were interested in this event was not that surprising, but these numbers suggested that there would be a point in looking at various country-specific Google domains in Europe as well.

Endogenous factors: language settings & IP address

As noted earlier, the geographic location appears to be one of – if not the most - influential cause of fluctuations in search rankings (Xing et al. 2014). These factors are probably some of the easiest to control for or, at least, influence to a large extent through the search settings and third-party software. Unfortunately, the case study did not take enough caution in managing the settings or changing the IP address. The assumption was that using the specific Google subdomains, google.at and google.dk, would be sufficient to get geo-specific results for Austria and Denmark, respectively. This did not turn out to be the case. Figure 2 shows the outcome of an attempt to query 'Felix Baumgartner' on google.at (the Austrian version of Google Search) to see the event from an Austrian perspective. Even though I specifically tried to avoid the particular Danish search results by using a different country-specific search domain (.at instead of .dk), Google overrides this. In the search results,

you find a video from the Danish-language version of Red bull's website (redbull.dk) as well as a news story from the largest Danish TV channel (nyhederne.tv2.dk). Furthermore, the language settings in the panel on the left side remain Danish. Since I did not change my IP address to a server in Austria and had Danish as my default language setting, it was quite likely these factors informed the search engine's decision to provide me with Danish search results on Google.at. Language settings could be fixed manually on the computer; but, as indicated in earlier studies of personalisation, the IP address also matters to a great extent for the geo-location factors.

[FIGURE 2 ABOUT HERE]

One obvious method to manipulate geo-location is to use some kind of IP scrambler that changes the IP address to an alternative IP address, e.g., through VPN servers, or disguises the IP address altogether, e.g., through The Onion Router (TOR) (Fernando, Du, and Ashman 2014). Many VPN providers allow users to specify from which country or region within a country they would like the IP address to be based. Thus, the geographic location can be directly manipulated and, thus, treated in a similar way to an independent variable in experimental designs. One can, for instance, alternate between IP addresses to influence the geo-location signals on which the search algorithm relies. Thus, it is theoretically possible for the same computer to appear as if it is based in different countries and, thus, retrieve search results that mimic those people in these countries would see. Another approach would, of course, be to recruit human participants physically based in the countries or regions under study. This is probably more desirable in many cases, since the whole project would not have to rely on a specific VPN provider, but it is resource demanding and has a risk that people will not comply or drop out.

To get a clearer idea of whether the language settings or IP addresses influence the constellation of search results, I conducted a small test (see Figure 3) some weeks after the study was concluded. Here, I queried 'Felix Baumgartner' on Google.de (German domain) with 4 different settings: One with the language set to Danish with an IP address in Copenhagen, Denmark (Figure 3a); one with the language set to German with an IP address in Denmark (Figure 3b); one with the language set to German and with a German IP address (Figure 3c); and one with the language set to Danish and with a German IP address (Figure 3d). The

greatest changes in the organic search results seem to come from the language settings. Notice, for example, how the country domains on Wikipedia follow the language settings and not the IP address. Meanwhile, the IP address informs the type of ads that are shown to the user in the top banners. It is simply not sufficient to change the IP address to direct Google toward the desired geo-location. Accordingly, the IP address seems, in fact, less important than language settings in influencing the ranking of organic search results.

[FIG3 ABOUT HERE]

Endogenous factors: web history, click behaviour, account logins

The personalisation of search results based on prior behaviour and other account signals (true personalisation, one could say) is very difficult to operate with and control for. Therefore, some have proposed various attempts to exclude these factors from influencing search rankings altogether. Richard Rogers, among others, has suggested operating with a 'research browser', that is, a browser cleansed of any history of prior usage that boots directly from scratch each time it is opened (Rogers 2013b). This browser would only be influenced by factors such as the keywords used, geo-location, and exogenous factors. This is probably a viable way to deal with personalisation issues if used in the stringent manner outlined by Rogers (Rogers 2013b). However, this method also introduces one major downside as I see it: it runs the risk of being too artificial and detached from real-world search situations. Even though people in general might be opposed to personalisation and targeted advertisement (Purcell, Brenner, and Rainie 2012), many people are either logged into Google when they search (knowingly or unknowingly), do not clear their search history and cookies, or do not think more generally about how their web behaviour influences the search engine. Accordingly, even though it might be possible to strip the search engine of some of the factors, it also comes with the cost of operating within an artificial environment. It might be a reliable way to mitigate some personalisation, but it is not necessarily valid.

In the case of the Red Bull Stratos, click behaviour was the only clear-cut indicator of a change in the rankings. At one point, I interacted with one of the links in the rankings, the Wikipedia page for Felix Baumgartner, and that resulted in a particular page attaining a much better ranking (from the bottom 3 sites to the top 4 sites) one hour later. Investigating web

history and account logins was not directly part of the design (the computers changed IP addresses and none of them logged into any accounts during the study). However, it would be possible to work with those factors as well. For account logins, one can alternate between being logged in and not (at least, with real-life participants) and compare differences. Web history would be more difficult to manipulate directly (for an attempt to do so with research profiles, consult Feuz, Fuller, and Stalder 2011). However, it is possible to assess the influence of individual browsing history (both web visits, previous searches, and so on) if all other factors are sought to be held constant. This would be a passive way of estimating influence. If there were no variations across participants with otherwise similar search profiles, then there would be a case for a limited influence of browsing history. This is a slightly problematic approach, since we cannot hold all other factors constant, because we do not know all the other factors, as discussed earlier. Nonetheless, it is not impossible to work with these factors in the design, either.

Exogenous factors: randomization & experimentation

Common among the endogenous factors is - as mentioned earlier - that the researcher can have a large degree of control on how they interfere with the project. The exogenous factors, on the other hand, pose a more imminent threat to consistent and reliable results both over time and with respect to participants. This is primarily because these factors are very difficult to control for. Experimentation and randomisation cannot be directly observed; but, in studies over periods of time, it is possible to observe their influence or lack thereof. If odd results cannot be reproduced across searchers or for the same searcher over time, then it should warrant caution in the interpretation of the results. It could be a sign that randomisation, experimentation or other exogenous factors are at work in the search algorithm. As long as we assume that randomisation and experimentation are randomly distributed across our searchers (research profiles or real-world participants) or, at least, not correlated with the endogenous factors we are interested in studying (e.g., geo-location), then it is actually a lesser problem for the validity of the study. Primarily, they serve as a caution for how many other factors can be involved in the engine and a reminder that we should always interpret search rankings very carefully due to their specific ontology as algorithmically generated content on the web.

Ways forward: experimental design or careful sampling

Instead of trying to explain all the factors that go into the algorithm or ignoring them altogether, researchers are advised to use the factors actively in the research design. This can be done by turning the focus to how search results for specific queries vary for individual searchers across space and over time. This means paying attention to variability in three dimensions: ***what*** is searched for (variations across search queries), ***who*** does the searching (variations across searchers), and ***when*** is the search conducted (variations over time). Thus, instead of treating these factors as a problem, we should seek to work with the variability they provide as an opportunity. To explore variations across search results the endogenous factors can be used actively as experimental variables (trying to hold some factors constant in the design of search profiles while letting others vary) or as criteria for sampling in the research design (e.g., by employing a "maximum variation sampling strategy"(Kuzel 1992) in which participants most different on one or more endogenous factors are selected for the study). Here, differences across searchers would indicate that the particular endogenous factors under study make a difference for the search rankings (e.g., the geo-location of participants), whereas similarity would indicate the opposite. Thus, such a design would be able to document and assess the variations in the sources to which people are exposed through Google Search in relation to news events and stories. This type of design lends itself to questions such as: which genre of web content (Helles 2013) is typically associated with different search queries? Which types of news brands are prominently displayed in the search results? Do rankings of genre and news brands change across individual searchers or over time – for instance, in relation to a major news event?

Depending on the emphasis, such a design can either be cross-sectional (variation across searchers *at one point in time*) or longitudinal (variation across *multiple points in time*) in some form or another. The short burst model discussed here is actually a variation of the two (it should ideally go across searchers and with multiple points within a short time frame). The short burst model documenting a pre-programmed news event is advantageous for research, because it allows for planning, setting up, and possible recruiting participants in advance.

However, this is, of course, not how most news stories play out. They are rather more sudden, disruptive, and either not scheduled in advance, such as natural disasters, or not made public before the act, such as terrorism (Katz and Liebes 2007). On these occasions, the cross-sectional model with research profiles seems most feasible initially to capture developments in real-time. Shortly thereafter, the design can be adapted to suit real-world participants or more carefully designed research profiles.

In general, whether to use research profiles programmed by the researcher or real-world participants depends on the study design. In general, the strength of using real-world participants is, first and foremost, that their search result would reflect more realistic search behaviour (in the best of all worlds, neither over- or underestimating the amount of privacy protection people use themselves). Research profiles, on the other hand, offer the researcher better control of the study design. Profiles could be set up with alternating characteristics (for instance, based on the design in Hannak et al. 2013 or Xing et al. 2014), which is easier to control experimentally but also risks becoming too detached from real-world users and thereby detrimental to the validity of the study. To achieve geographical variation in research profiles, which the literature so far has identified as the most important factor causing personalisation, one would have to take great caution in selecting proper ways of masking IP addresses and change language settings accordingly.

A final issue that has not been treated directly so far is how to report the search results in research. As should be clear from the discussion so far, it is very important how the data has been generated (e.g., as screen dumps) for what one can conclude from the data. At a minimum, time stamps should be supplied for the exact moment the query was posted on Google; the language setting and geo-location (e.g., as IP address) should be included, and it should be clearly specified on which device (computer, tablet, mobile) the search was conducted. It should also be noted whether factors were specifically manipulated (e.g., whether one was logged into Google services). Lastly, it should be clear from the context whether the data were generated by a human searcher or through a program. If the latter is the case, it should be made clear how this program might distort the factors (e.g., personalisation factors generally). In short, all the information that could theoretically be thought to make a difference should be included either on the individual screen dumps or generally for the context of the study.

On a final note, Google search is not the only important algorithmically generated material to document. The algorithmic curation of content plays a growing importance online. Not only are search engines and social media, such as Facebook (Bucher 2012) and YouTube (Dijck 2013), using algorithms to display and rank content to the users but, to an increasing extent, so are websites more generally (Mayer-Schönberger and Cukier 2013) and news sites in particular (Thurman and Schifferes 2012). Personalisation of content and segmentation of web traffic into targetable user groups is a valuable business for web companies and, very likely, a practice we will see spread to even more websites and mobile apps in the coming years (Couldry and Turow 2014). The methodological challenges discussed here are relevant to broader studies of personalisation on the internet.

Conclusion

Search engines provide a window into the changing association between websites and keywords across cultures and countries and over time. As such, they offer journalism and news researchers an opportunity to study how search engines, in this case Google, mediate news events and stories online. Here, I have discussed how the ontological nature of search results as peculiar documents on the web condition the way we are able to archive them as sources for further research. The most prominent factors informing the search engine algorithms have been identified and assessed. Through a small case study, ways in which researchers can operate with the influence of endogenous factors (keywords, language settings, geo-location, web history, and click behaviour) and consider the effects of the exogenous factors (experimentation, randomization) have been proposed. In the last part, a new approach to studying search results has been suggested, which builds on purposeful sampling of real-world participants or constructed profiles suited for more qualitative studies. This is both an analytically fruitful approach, which offers new perspectives on search engine research, and one that actively works with - instead of against – the algorithm. Finally, perspectives on studying algorithms in a broader context than search have been offered.

References

Agarwal, Amit. 2015. “How to Scrape Google Search Results Inside a Google Sheet.” *Digital Inspiration*, June 4. http://www.labnol.org/internet/google-web-scraping/28450/.

Borra, Erik, and René König. 2013. “Googling 9/11: The Perspectives of a Search Engine on a Global Event.” In *Society of the Query* #2. Amsterdam: Institute of Network Cultures.

Bozdag, Engin. 2013. “Bias in Algorithmic Filtering and Personalization.” *Ethics and Information Technology* 15: 209–27.

Brin, Sergey, and Lawrence Page. 1998. "The anatomy of a large-scale hypertextual Web search engine." *Computer networks and ISDN systems* 30(1): 107-117.

Bro, Peter, and Filip Wallberg. 2014. “Digital Gatekeeping.” *Digital Journalism* 2 (3): 446–54.

Bruns, Axel, Jean E. Burgess, Kate Crawford, and Frances Shaw. 2012. “#qldfloods and @QPSMedia: Crisis Communication on Twitter in the 2011 South East Queensland Floods.” *ARC Centre of Excellence for Creative Industries and Innovation*. http://eprints.qut.edu.au/48241/1/floodsreport.pdf.

Brügger, Niels. 2011. “Web Archiving – Between Past, Present, and Future.” In *The Handbook of Internet Studies*, edited by Mia Consalvo and Charles Ess, 24–42. Chichester: Wiley-Blackwell.

Bucher, T. 2012. “Want to Be on the Top? Algorithmic Power and the Threat of Invisibility on Facebook.” *New Media & Society* 14 (7): 1164–80.

Couldry, Nick, and Joseph Turow. 2014. “Big Data, Big Questions| Advertising, Big Data and the Clearance of the Public Realm: Marketers’ New Approaches to the Content Subsidy.” *International Journal of Communication* 8: http://ijoc.org/index.php/ijoc/article/view/2166.

Dijck, José van. 2013. *The Culture of Connectivity: A Critical History of Social Media*. Oxford & New York, NY: Oxford University Press.

Edelman, Benjamin. 2010. “Hard-Coding Bias in Google ‘ Algorithmic ’ Search Results.” *Benjamin Edelman*, November 15. http://www.benedelman.org/hardcoding/.

Edelman, Benjamin, and Benjamin Lockwood. 2011. “Measuring Bias in ‘Organic’ Web Search.” *Benjamin Edelman*, January 19. http://www.benedelman.org, http://www.benedelman.org/searchbias/.

Fernando, Anisha T. J., Jia Tina Du, and Helen Ashman. 2014. "Personalisation of Web Search: Exploring Search Query Parameters and User Information Privacy Implications-The Case of Google." In *Proceeding of the 1st International Workshop on Privacy-Preserving IR: When Information Retrieval Meets Privacy and Security* (PIR 2014), 31–36.

Feuz, Martin, Matthew Fuller, and Felix Stalder. 2011. "Personal Web Searching in the Age of Semantic Capitalism: Diagnosing the Mechanisms of Personalisation." *First Monday* 16 (2). http://firstmonday.org/ojs/index.php/fm/article/view/3344/2766.

Granka, Laura A. 2010. "The Politics of Search: A Decade Retrospective." *The Information Society* 26 (5): 364–74.

Hannak, Aniko, Piotr Sapiezynski, Arash Molavi Kakhki, Balachander Krishnamurthy, David Lazer, Alan Mislove, and Christo Wilson. 2013. "Measuring Personalization of Web Search." In *Proceedings of the 22Nd International Conference on World Wide Web* (WWW '13), 527–38.

Heitner, Darren. 2012. "Red Bull Stratos Worth Tens of Millions of Dollars in Global Exposure for The Red Bull Brand." *Forbes*, October 15. http://www.forbes.com/sites/darrenheitner/2012/10/15/red-bull-stratos-worth-tens-of-millions-of-dollars-in-global-exposure-for-the-red-bull-brand/.

Helles, Rasmus. 2013. "The Big Head and the Long Tail: An Illustration of Explanatory Strategies for Big Data Internet Studies." *First Monday* 18 (10): http://firstmonday.org/ojs/index.php/fm/article/view/4874/3753.

Hillis, Ken, Michael Petit, and Kylie Jarrett. 2013. *Google and the Culture of Search*. New York, NY & Oxon, UK: Routledge.

Jensen, Klaus Bruhn. 2010. "New Media, Old Methods–Internet Methodologies and the Online/Offline Divide." In *The Handbook of Internet Studies*, edited by Mia Consalvo and Charles Ess, 43-58. Chichester: Wiley-Blackwell.

Jiang, Min. 2012. "The Business and Politics of Search Engines: A Comparative Study of Baidu and Google's Search Results of Internet Events in China." *New Media & Society* 16 (2): 212–33.

Karlsson, Michael, and Jesper Strömbäck. 2010. "Freezing The Flow Of Online News." *Journalism Studies* 11 (1): 2–19.

Karpf, David. 2012. "Social Science Research Methods in Internet Time." *Information, Communication & Society* 15 (5): 639–61.

Katz, Elihu, and Tamar Liebes. 2007. "'No More Peace!': How Disaster, Terror and War Have Upstaged Media Events." *International Journal of Communication* 1: 157–66. http://ijoc.org/index.php/ijoc/article/view/44.

Kautsky, Robert, and Andreas Widholm. 2008. "Online Methodology: Analysing News Flows of Online Journalism." *Westminster Papers in Communication and Culture* 5 (2): 81–97.

Kuzel, Anton J. 1992. "Sampling in Qualitative Inquiry." In *Doing Qualitative Research*, edited by B. F. Crabtree and W. L. Miller, 31–44. Thousand Oaks, CA: Sage Publications.

Lewandowski, Dirk. 2005. "Web Searching, Search Engines and Information Retrieval." *Information Services & Use* 25: 137–47.

Makino, Takaki, Chaesang Jung, and Doantam Phan. 2015. "Finding More Mobile-Friendly Search Results." *Google - Webmaster Central Blog,* February 26. http://googlewebmastercentral.blogspot.dk/2015/02/finding-more-mobile-friendly-search.html.

Marres, Noortje, and Esther Weltevrede. 2013. "Scraping the Social? Issues in Real-Time Social Research." *Journal of Cultural Economy* 6 (3): 313–35.

Mayer-Schönberger, Viktor, and Kenneth Cukier. 2013. *Big Data: A Revolution that Will Transform How We Live, Work, and Think*. New York, NY: Houghton Mifflin Harcourt.

Mitchell, Amy, Mark Jurkowitz, and Kenneth Olmstead. 2014. "How Readers Get to News Sites: Social, Search and Direct - Pathways to Digital News." *Pew Research Center - Journalism & Media*, March 13. http://www.journalism.org/2014/03/13/social-search-direct/.

Pan, Bing, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka. 2007. "In Google We Trust: Users' Decisions on Rank, Position, and Relevance." *Journal of Computer-Mediated Communication* 12 (3): 801–23

Pariser, Eli. 2012. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. New York, NY: Penguin Books.

Purcell, Kristen, Joanna Brenner, and Lee Rainie. 2012. "Search Engine Use 2012". *Pew Research Center – Internet, Science & Tech,* March 9. http://www.pewinternet.org/2012/03/09/main-findings-11/.

Rogers, Richard. 2013a. "Plenary Session: 'The Network Tradition in Communication Research and Scholarship.'" In *International Communication Association (ICA) Conference*, London (June 17-22, 2013). London.

———. 2013b. *Digital Methods*. Cambrdige, Mass.: MIT Press.

Sanz, E., and J. Stancik. 2013. "Your Search - 'Ontological Security' - Matched 111,000 Documents: An Empirical Substantiation of the Cultural Dimension of Online Search." *New Media & Society* 16(2): 252-270.

Spink, Amanda, Yashmeet Khopkar, Prital Shah, and Sandip Debnath. 2003. "Search Engine Personalization: An Exploratory Study." *First Monday* 8(7): http://firstmonday.org/ojs/index.php/fm/article/view/1063.

Thurman, Neil, and Steve Schifferes. 2012. "The Future of Personalization at News Websites." *Journalism Studies* 13 (5-6): 775–90.

Vis, Farida. 2013. "Twitter as a Reporting Tool for Breaking News." *Digital Journalism* 1 (1): 27–47.

Xing, Xinyu, Wei Meng, Dan Doozan, Nick Feamster, Wenke Lee, and AlexC. Snoeren. 2014. "Exposing Inconsistent Web Search Results with Bobble." In *Passive and Active Measurement SE - 13*, edited by Michalis Faloutsos and Aleksandar Kuzmanovic, 8362:131–40.

Zimmer, Michael. 2008. "The Externalities of Search 2.0: The Emerging Privacy Threats When the Drive for the Perfect Search Engine Meets Web 2.0." *First Monday* 13 (3): http://firstmonday.org/ojs/index.php/fm/article/view/2136

Zuckerman, Ethan. 2011. "In Soviet Russia, Google Researches You! | ... My Heart's in Accra." ...My Heart's in Accra. http://www.ethanzuckerman.com/blog/2011/03/24/in-soviet-russia-google-researches-you/.

---

[i] According to a search for "david cameron" on Google Trends: https://www.google.com/trends/explore#q=david%20cameron%2C%20%2Fm%2F04qdjv&cmpt=q&tz= (accessed 12 May 2015).

[ii] See, for instance, the official remarks from Google on https://developers.google.com/custom-search/ (accessed 16 January 2015).

[iii] Googles Terms of Service (ToS): http://www.google.com/intl/en/policies/terms/ (accessed 20 August 2015).

[iv] The Danish web archive (netarkivet.dk) archives certain culturally and politically important websites on a routine basis and then archives an extensive number of websites determined on an *ad hoc* basis for specific pre-planned or suddenly occurring events. In that way, they combine the models presented here.

[v] Comparison made with the free tool Google Trends. Available at: www.google.com/trends/ .